(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 September 2001 (20.09.2001)

PCT

(10) International Publication Number
WO 01/68807 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US01/08590

(22) International Filing Date: 16 March 2001 (16.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/190,362 16 March 2000 (16.03.2000) US
60/272,847 1 March 2001 (01.03.2001) US

(71) Applicant *(for all designated States except US)*: **FRED HUTCHINSON CANCER RESEARCH CENTER** [US/US]; Office of Technology Transfer, 1100 Fairview Avenue North, M/S: C2M 027, Seattle, WA 98109-1024 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **VAN STEENSEL, Bas** [NL/NL]; Nieuwegrachtje 1-3, NL-1011 VP Amsterdam (NL). **HENIKOFF, Steven** [US/US]; 4711 51st Place SW, Seattle, WA 98116 (US).

(74) Agents: **POOR, Brian, W.** et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111 (US).

(81) Designated States *(national)*: AU, CA, JP, US.

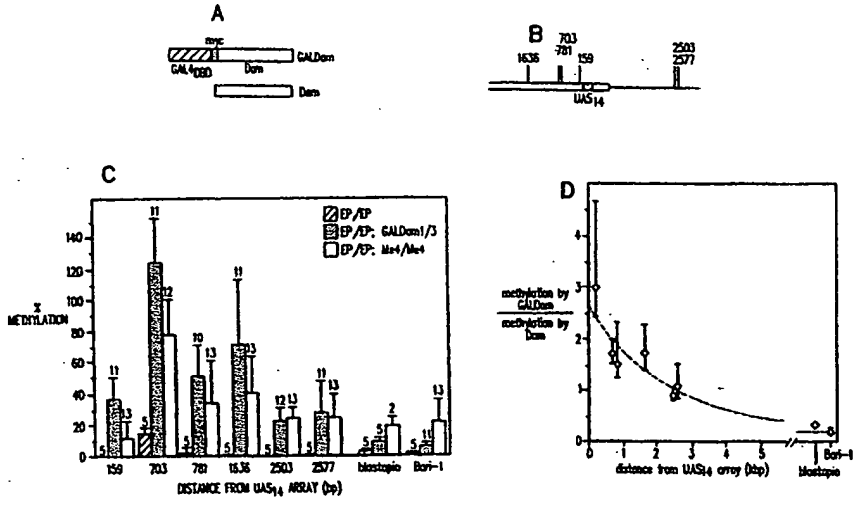
(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *without international search report and to be republished upon receipt of that report*

(48) Date of publication of this corrected version:
27 December 2001

[Continued on next page]

(54) Title: IDENTIFICATION OF *IN VIVO* DNA BINDING LOCI OF CHROMATIN PROTEINS USING A TETHERED NUCLEOTIDE MODIFICATION ENZYME



(57) Abstract: A novel technique is provided, designated DamID, for the identification of DNA loci that interact *in vivo* with specific nuclear proteins in eukaryotes. By tethering a DNA modification enzyme, in particular, *E. coli* DNA adenine methyl transferase (Dam), to a chromatin protein. The DNA modification enzyme (Dam) can be targeted *in vivo* to the native binding loci of the protein, resulting in local DNA modification. Sites of DNA modification can subsequently be mapped using modification-specific restriction enzymes, antibodies, or DNA array methods. DNA Modification Identification (DamID) has potential for genome-wide mapping of *in vivo* target binding sites of chromatin proteins in various eukaryotes.

WO 01/68807 A2

binding motif

<400> 6
tgagagagc 9

SEQUENCE LISTING

<110> FRED HUTCHINSON CANCER RESEARCH CENTER
van Steensel, Bas
Henikoff, Steven

<120> IDENTIFICATION OF IN VIVO DNA BINDING SITES OF CHROMATIN PROTEINS USING A TETHERED DNA MODIFICATION ENZYME

<130> 14538A-62-3PC

<140> PCT/US01/
<141> 2001-03-16

<150> 60/
<151> 2001-03-01

<150> 60/190,362
<151> 2000-03-16

<160> 6

<170> PatentIn Ver. 2.1

<210> 1
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: myc-epitope tag

<400> 1
Glu Gln Lys Ile Ser Glu Glu Asp Leu
1 5

<210> 2
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: PCR probe

<400> 2

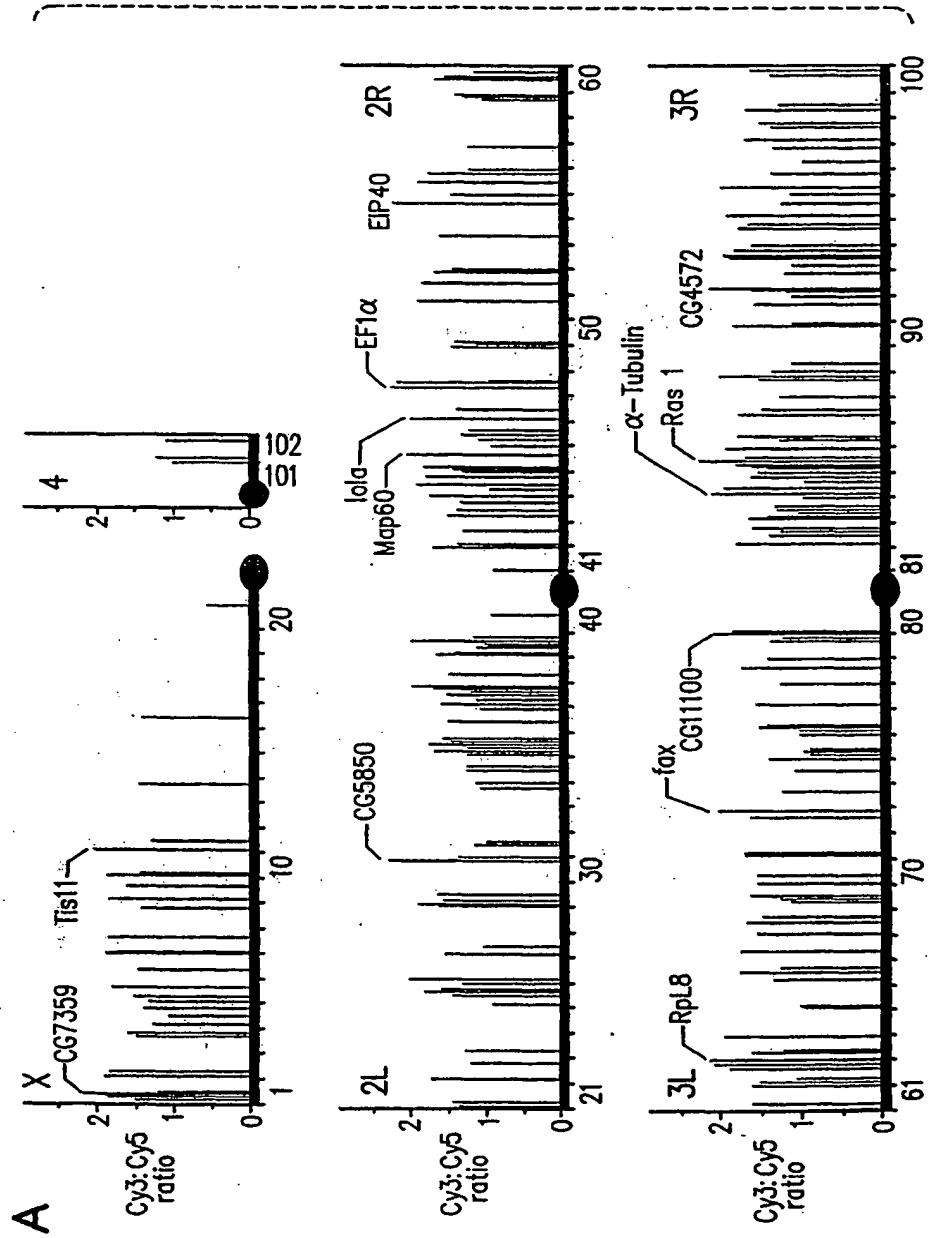


FIG. 6A

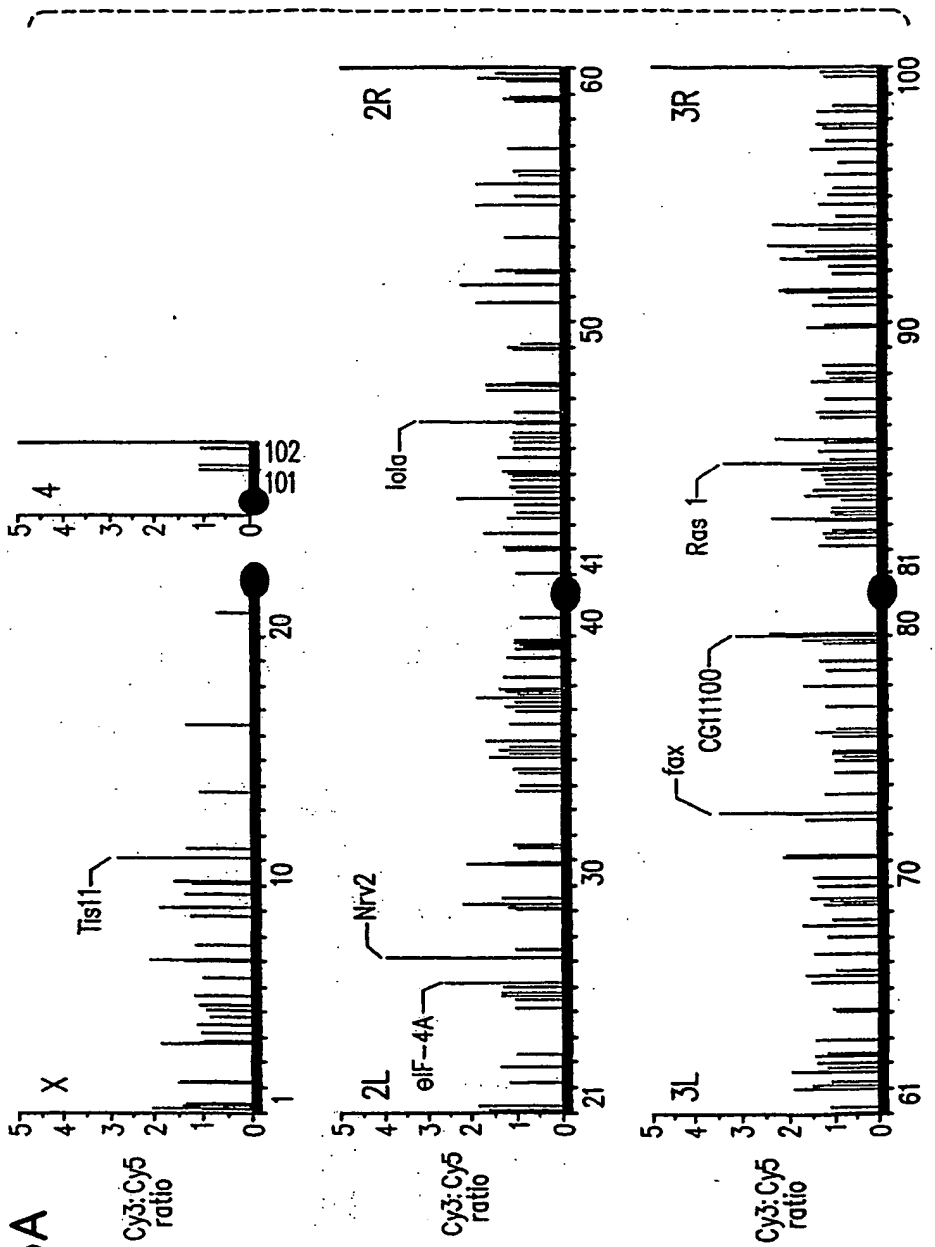

X
2L
2R
3L
3R
Cy3:Cy5 ratio
Tis11
elF-4A
Nrv2
Iola
fax
CG11100
Ras 1
102
101

FIG. 5A

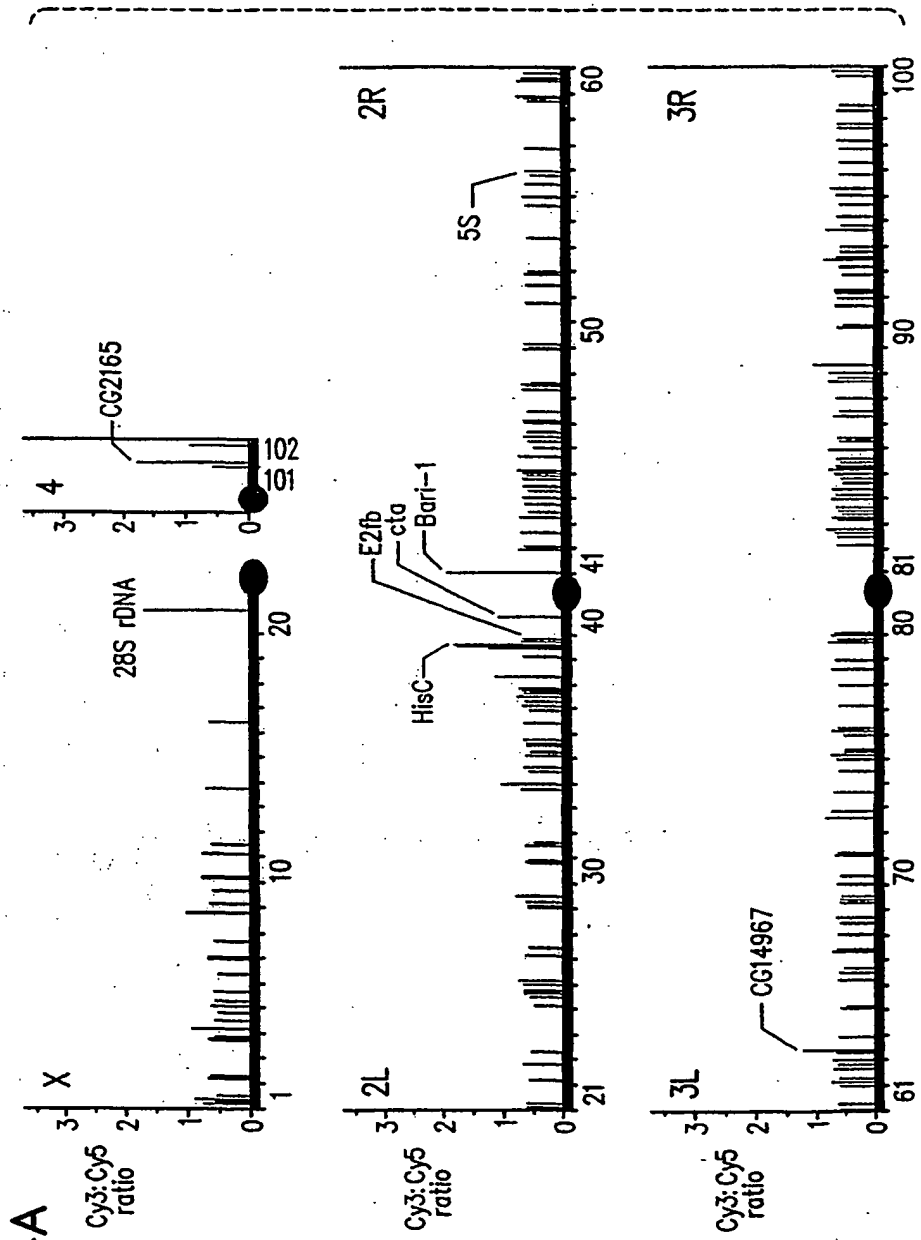

X
2L
2R
3L
3R
4
Cy3:Cy5 ratio
28S rDNA
CG2165
102
101
E2fb
cta
Bari-1
HisC
5S
CG14967

FIG. 4A

FIG. 1D

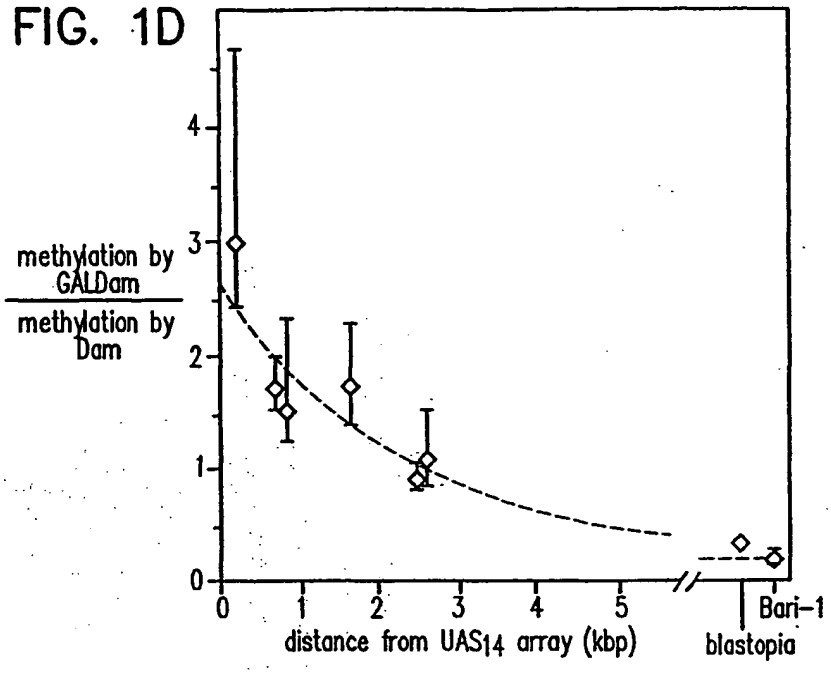


FIG. 2

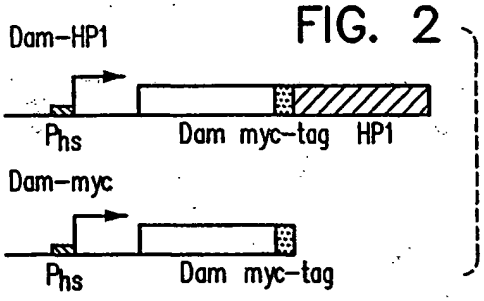
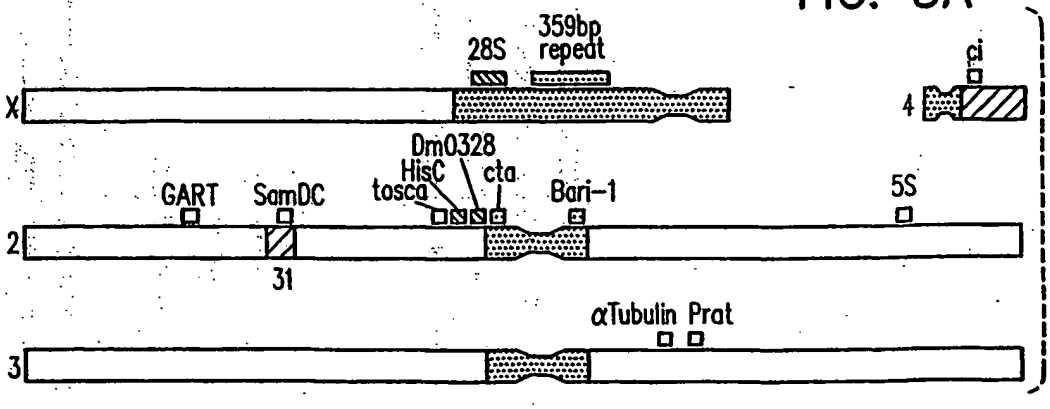


FIG. 3A



28. The method of claim 27, wherein the label is chemiluminescent, an enzyme, a fluorophor, or a radioactive moiety.

29. The method of claim 28, wherein the fluorescent label is fluorescein, phycoerythrin (PE), Cy3, Cy5, Cy7, Texas Red, allophycocyanin (APC), Cy7APC, Cascade Blue, or Cascade Yellow.

30. The method of claim 25, wherein the bound antibody is detected by a labeled second antibody.

31. The method of claim 24, wherein the array comprises DNA, cDNA, DNA comprising substantially only chromatin binding regions, RNA, or RNA comprising substantially only protein binding regions.

10. The method of claim 9, wherein the antibody is polyclonal, monoclonal, a single chain antibody, a chimeric antibody, or an antigen binding fragment thereof.

11. The method of claim 9, wherein the antibody is labeled.

12. The method of claim 11, wherein the label is chemiluminescent, an enzyme, a fluorophor, or a radioactive moiety.

13. The method of claim 12, wherein the fluorescent label is fluorescein, phycoerythrin (PE), Cy3, Cy5, Cy7, Texas Red, allophycocyanin (APC), Cy7APC, Cascade Blue, or Cascade Yellow.

14. The method of claim 9, wherein the bound antibody is detected by a labeled second antibody.

15. The method of claim 8, wherein the array comprises DNA, cDNA, DNA comprising substantially only chromatin binding regions, RNA, or RNA comprising substantially only protein binding regions.

16. A method for producing a profile of chromatin protein loci for a cell population of interest comprising;

transfecting the cell population with a plurality of expression vectors capable of expressing a plurality of chromatin protein-nucleotide modification enzyme fusion proteins, each expression vector comprising a nucleic acid encoding a low efficiency promoter operativelty associated with a nucleic acid encoding the chromatin proteins and a nucleic acid encoding a nucleotide modification enzyme;

culturing the transfected cells for a period of time sufficient for expression of and binding of each of the plurality of chromatin protein-nucleotide modification enzyme fusion proteins; and

detecting the loci for each of the nucleotide modifications within the chromatin of the cell population; therefrom determining the profile of chromatin protein loci for the cell population.

Dam ($p<0.003$), but not for Dam-GAF ($p=0.25$). The somewhat higher noise level in the Dam-GAF data obtained in the assays may preclude detection of exclusion from HP1 binding sites. Finally, comparison of GAF and DmSir2-1 revealed a subset of genes that were associated with both proteins. Biochemical analysis may reveal whether the two proteins can be part of one protein complex, or whether they bound separately to different regions in the same genes.

To confirm the relative distributions of the three proteins their immunocytochemical staining patterns were examined. As provided above in Example 1, HP1 in Kc cells was associated with a large chromocenter. In contrast, the DmSir2-1-Dam fusion protein appeared to be associated with the euchromatic compartment, and was essentially excluded from HP1-containing regions. Likewise, the GAF-Dam fusion protein was located in the euchromatin compartment and mostly absent from the chromocenter. These cytological results were in agreement with the molecular mapping data, and confirmed that GAF and DmSir2-1 were preferentially associated with non-heterochromatic regions.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be obvious that certain changes and modifications may be practiced within the scope of the appended claims. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

N-terminus of GAF. Again, results were reproducible, with r in pairwise comparisons of three experiments ranging from 0.81-0.93. Importantly, the C- and N-terminal fusion proteins gave similar results, r = 0.80, in two independent comparisons), although Cy3:Cy5 ratios with Dam-GAF cover a smaller dynamic range than with GAF-Dam. These results strongly suggest that Dam, when fused to either end of GAF, does not interfere with correct targeting of GAF.

Genes that appear to strongly bind GAF have no common function or expression pattern. Because *in vitro* binding assays and *in vivo* cross-linking studies have shown that GAF binds GA-rich regulatory elements (Biggin et al., *Cell* 53:699-711(1988); Soeller et al., *Mol. Cell. Biol.* 13:7961-7970 (1993); Strutt et al., *EMBO J.* 16:3621-3632 (1997); O'Brien et al., *Genes Dev.* 9:1098-1110(1995)), the GAF target loci identified by the mapping were investigated to determine whether they were enriched in such elements. Indeed, loci that display moderate to strong GAF binding have significantly higher average densities of GAGAG (SEQ ID NO: 4) and GAGAGAG (SEQ ID NO: 5) sequences than loci with low GAF binding (Fig. 5B) providing strong evidence that bona fide target loci of GAF were identified.

Target loci of DmSir2-1: Finally, target loci of a Drosophila homolog of budding yeast Sir2 were mapped by the methods of the present invention. In *S. cerevisiae*, Sir2 plays a role in silencing of genes in the silent mating-type loci, telomeric regions, and the rDNA locus (Guarente, *Genes Dev.* 14:1021-1026 (2000); Gartenberg, *Curr. Opin. Microbiol.* 3:132-137 (2000)). In Drosophila, five Sir2-like proteins have been predicted by sequence analysis (Frye, *Biochem. Biophys. Res. Commun.* 273:793-798 (2000)). Of these five, the Sir2-like protein that was found to be most closely related to *S. cervisiae* Sir2 was chosen. The selected protein has been referred to herein as DmSir2-1. The homology to yeast Sir2 suggested that DmSir2-1 might be associated with heterochromatin in Drosophila, but no experimental studies of DmSir2-1 have been reported.

Mapping results obtained with a DmSir2-1-Dam fusion protein are shown in Fig. 6. DmSir2-1 demonstrated association with numerous genes in a reproducible fashion (r=0.81 between two independent experiments). Among the strongest DmSir2-1 binding loci were several euchromatic, constitutively expressed genes such as genes encoding translation factors, putative ribosomal proteins, α-tubulin, hsc4 and EIP40. This suggests that DmSir2-1 binds to active genes, unlike yeast Sir2.

Biol. 19:4366-4378 (1999)). A small number of pericentric target loci of HP1 have been identified (van Steensel and Henikoff, *Nature Biotechnol.* 18:424-428 (2000)), but the nature of the HP1 binding sites on the euchromatic arms is unknown.

A scatter diagram of the hybridization signals measured for Cy3 (Dam-HP1) vs Cy5 (Dam) showed that the majority of cDNAs display an almost identical Cy3:Cy5 ratio (*i.e.*, were located on a single diagonal in the scatter diagram), indicating no detectable association of the corresponding genes with HP1. However, a distinct set of cDNAs demonstrated a clear offset from this diagonal towards higher Cy3:Cy5 ratios. These cDNAs must represent target loci of HP1. The absence of data points with lower Cy3:Cy5 ratios demonstrated that tethering of Dam to HP1 caused an increase in methylation of HP1 target loci, but not a decrease in methylation levels of non-target loci.

The probed loci are represented in Fig. 4A on the standard polytene chromosome map, showing their relative HP1 binding (*i.e.*, Cy3:Cy5 ratios). Most loci display a constant Cy3:Cy5 ratio (approximately 0.5-0.6), which was interpreted as non-targeted 'background' methylation. However, several loci demonstrated a considerably higher ratio, implying HP1 binding. Although the cutoff between 'target' and 'non-target' Cy3:Cy5 ratios was arbitrary, it is important to note that the differences in Cy3:Cy5 ratios between probed loci were highly reproducible. Pair-wise comparisons of three independent experiments showed correlation coefficients between 0.95 and 0.99. Hence, loci that demonstrated only a mild increase in Cy3:Cy5 ratio over background levels (*e.g.*, gene CG14967, Fig. 4A) were likely to be associated with HP1 *in vivo*, although the local HP1 concentration may be lower than at other target loci with higher Cy3:Cy5 ratios. Moreover, differences in Cy3:Cy5 ratios between genes in the present assay may somewhat underestimate differences in protein binding. By Southern blot analysis it was found that the Ban-1 locus displays about 8-fold higher HP1-targeted methylation than the 5S rDNA locus (Example 1), yet the present microarray analysis indicated only about a 4-fold difference. Such a microarray-specific compression effect has been observed previously (Pollack et al., *Nature Genet.* 23:41-46 (1999)).

Among the target loci of HP1 detected were genes located near pericentric heterochromatin, or on the largely heterochromatic chromosome 4. Both the histone gene cluster (HisC) and the *cta* gene, located near the centromere on the left arm of chromosome 2, were found to be associated with HP1, in agreement with previous observations provided above in Example 1. In contrast, the *Ef*2b gene, which lies between these two loci, showed no detectable HP1 binding, suggesting a discontinuous distribution of HP1 in this region.

citrate, pH 7.0) onto poly-lysine coated microscope slides using an OmniGrid high-precision robotic gridder (GeneMachines, San Carlo, CA).

One μg of purified methylated DNA was labeled with Cy3- or Cy5-dCTP (Amersham Pharmacia, Piscataway, NJ) by random priming (Pollack et al., *Nature Genet.* 23:41-46 (1999)). Labeled experimental and reference DNA samples were mixed and hybridized to a microarray in 3X SSC in the presence of 20 μg poly [dA.dT], 100 μg yeast tRNA, and 25 μg unlabeled *Dpn*I-digested plasmid encoding the fusion protein that was used for transfection. Hybridization was performed at 63 °C for 16 hours followed by sequential washings in 1X SSC, 0.03% SDS, 1X SSC, 0.2X SSC, and 0.05X SSC. Washed arrays were spun dry in a centrifuge and immediately scanned using a GenePix 4000 fluorescent scanner (Axon Instruments, Inc., Foster City, CA).

Data analysis: Image processing was performed using GenePix 3.0 image analysis software. Statistical analysis was performed using StatView software (Abacus Concepts, Berkeley, CA). Cy3:Cy5 ratios were normalized using Drosophila total genomic DNA (spotted 16 times on each microarray slide) as an internal standard. Thus, a Cy3:Cy5 value of 1 represents the average level of binding of the chromatin protein along the entire genome.

Five ESTs initially thought to represent unique genes along the euchromatic arms (based on the available 5' sequence) were identified as HP1 targets. Sequencing of both 5' and 3' ends indicated that these clones are hybrids of a unique gene and a repetitive sequence. It was presume that this was a cloning artifact that occurred during the construction of the CK library (Kopczynski et al., *Proc. Natl. Acad. Sci. USA* 95:9973-9978 (1998)). These clones are represented in the Dispersed Elements section of Figures 4 through 6. The fact that these chimeric clones were identified as HP1 targets underscores the sensitivity of the assay of the present invention.

Analysis of the density of $(GA)_n$ elements was carried out as follows. Fifteen loci showing low GAF-Dam binding (Cy3:Cy5 ratios 0.97 ± 0.09) and 15 loci showing high GAF-Dam binding (Cy3:Cy5 ratios 2.77 ± 0.65) were selected based on availability of the complete probe sequence. Corresponding genomic sequences were obtained from the BDGP/Celera genomic database and covered the region encoding the cDNA fragment present on the microarray. Any introns smaller than 5 kb, as well as 3 kb of sequence upstream and downstream of the probed region were included in the analysis, because methylation by tethered Dam extends in cis over a few kbs. On average, about 7.5 kb was analyzed per probed region. GAGAG (SEQ ID NO: 4) and GAGAGAG (SEQ ID NO: 5)

HisC probe 3 (R)	X14215	2490, 3835
STS Dm0328	BDGP[c]	135, 393
28S (R)	M21017	5018, 5492
SamDC	Y11216	410, 961
ci	U66884	12262, 12865

[a] (R) indicates that the probed region is predominantly present as a tandem repeat array
[b] cDNA sequence; the corresponding genomic sequence (GenBank accession AF211849) contains two approximately 50 bp introns
[c] Berkeley Drosophila Genome Project database

EXAMPLE 2

The present example provides large-scale mapping of *in vivo* binding sites of chromatin proteins, using a combination of targeted DNA modification and microarray detection. Three distinct chromatin proteins in Drosophila Kc cells were mapped and each were found to associate with specific sets of genes. HP1 was found, as above, binds predominantly to pericentric genes and transposable elements, GAGA factor associates with euchromatic genes that are enriched in $(GA)_n$ motifs. Surprisingly, a Drosophila homolog of yeast *Sir2* was found to associate with several active genes and was excluded from heterochromatin.

The materials and methods used for microarray detection of modified DNA regions peripheral to the binding loci of the DNA binding protein were as follows.

Plasmids: Vectors for expression of myc-tagged Dam and Dam-HP 1 were described above. A cDNA encoding full-length DmSir2-1 (GenBank accession AF068758) was obtained by PCR amplification from a Drosophila ovary cDNA library and cloned into pCMycDam as above, resulting in plasmid pSir2α-MDam. Sequencing of the cloned PCR product revealed that six nucleotides encoding Phe_{289} and Gln_{290} were missing. The same polymorphisms were also present in a genomic sequence (Genbank AE003639). Dam was fused to the C-terminus of DmSir2-1 because it had previously been found that addition of Green Fluorescent Protein to this end of yeast Sir2 did not interfere with correct subnuclear targeting of Sir2 (Cuperus et al., *EMBO J.* 19:2641-2651 (2000)).

Full length GAF (the 519 amino acid isoform (Benyajati et al., *Nucleic Acids Res.* 25:3345-3353 (1997), incorporated herein by reference) was amplified from plasmid

chromosome 2, has been speculated to have heterochromatic features (Fitch et al., *Chromosoma* 99:118-124 (1990)). The rDNA repeat has been mapped to the heterochromatic part of the X chromosome (Hilliker et al., *Cell* 21:607-619 (1980)), yet during interphase it is packaged inside the transcriptionally active nucleolus (Scheer et al., *Opin. Cell. Biol.* 11:385-390 (1999)). Somewhat surprisingly, it was found that both loci (three different regions in HisC, and the 28S gene in the rDNA repeat) displayed Dam-HP1/Dam methylation ratios that were intermediate between euchromatin and heterochromatin (Fig. 3C). Since both loci are tandem repeats, it is possible that only a fraction of the repeats was associated with HP1. Alternatively, the association of HP1 with these loci may be cell cycle regulated. A similar level of Dam-HP1 targeted methylation was observed for sequence tag STS Dm0328, which is located in the banded region proximal to HisC. Possibly, HP1 'spreads' from pericentric heterochromatin into the flanking euchromatin to include HisC.

Finally, the *cubitus interruptus (ci)* and *S-adenosyl decarboxylase (SamDC)* genes were tested. These genes are located in the banded part of chromosome 4 and in region 31 on chromosome 2, respectively. Both regions are decorated by antibodies against HP1 (James et al., *Eur. J. Cell. Biol.* 50:170-180 (1989)). *ci* showed levels of HP1-targeted methylation that were lower than in heterochromatic loci, but significantly higher than in euchromatic loci (Fig. 3B and Fig. 3C), indicating that HP1 is associated with this gene. In contrast, *SamDC* showed low levels of targeted methylation, suggesting that this gene was not abundantly associated with HP1. A detailed map of HP1 associations can be obtained in the future by systematic analysis of a large number of sequences throughout the genome.

Discussion

The data provided herein demonstrate that DamID can be used to identify sequences that interact *in vivo* with specific proteins. Targeting of Dam leads to up to an approximately 10-fold enrichment of methylation in the vicinity of binding sites of the Dam fusion partner, which is sufficient for positive identification of most target sequences, and for detecting quantitative differences in protein-target interactions. The 'background' methylation throughout the genome by Dam fusion proteins was attributed to the intrinsic DNA binding activity of Dam, which would compete with the sequence- or locus-specific interactions of its fusion partner. However, it is also possible that chromatin proteins are rather promiscuous in their interactions *in vivo*.

euchromatic loci. It has been thought that HP1 is recruited to specific genomic loci by other chromatin proteins (Platero et al., *EMBO J.* 14:3977-3986 (1995)). To identify HP1 target loci, a myc-epitope tagged Dam-HP1 fusion protein construct driven by a heat-shock promoter (Fig. 2) was transfected into Drosophila Kc cells, and the resulting methylation patterns were mapped. As a control. myc-tagged Dam (Dam-myc) was expressed.

In order to demonstrate that fusion to Dam did not impair correct targeting of HP1, advantage was taken of the observation that HP1 in Kc cells was predominantly located in a large discrete compartment in the nucleus. This compartment represented clustered pericentric heterochromatin of all chromosomes, since all centromeres were generally located within this compartment. In agreement with this, AT-rich heterochromatic satellite repeats, which were visible as DAPI-bright regions, were closely associated with the HP1 compartment. HP1 was also seen in a few small brightly stained dots scattered throughout the nucleoplasm.

After heat shock induction, the Dam-HP1 fusion protein (detected with an antibody against the myc epitope) showed a subnuclear distribution pattern that strongly resembled that of endogenous HP1. Both the large compartment, closely associated with the DAPI-bright regions, and a few small dots scattered throughout the nucleus were observed. This indicated that the fusion protein was correctly targeted to natural HP1 binding sites. In contrast, after heat-shock induction the Dam-myc protein showed a very weak staining throughout the cell, with no indication of subnuclear targeting. In the absence of heat shock induction the Dam-HP1 or Dam-myc proteins were not detectable by immunofluorescence, indicating very low expression levels under those conditions.

Whether expression of Dam-HP1 leads to preferential methylation of heterochromatic DNA was also tested. Sites of methylation were visualized *in situ* using a rabbit antiserum (Bringmann et al., *FEBS Lett.* 213:309-315 (1987)) against methyl-N^6-adenine (^{m6}A). After heat-shock induction, cells transfected with either Dam or Dam-HP1 showed strong ^{m6}A staining throughout the nucleus. Cells transfected with an empty vector showed no nuclear staining. Strikingly, in the absence of heat shock induction, cells transfected with Dam-HP1 generally showed staining of the large, discrete nuclear compartment associated with DAPI-bright regions. Co-staining of ^{m6}A with an antibody against endogenous HP1 confirmed that methylation was mostly restricted to the heterochromatic compartment. An imprecise correspondence of the two staining patterns was expected because of a lack of GATCs in the simple satellites in *Drosophila* heterochromatin. In metaphase chromosomes of cells transfected with Dam-HP1, ^{m6}A was

the resulting DNA preparation was incubated 16 hours at 37 °C with or without 2 units *Dpn*II. After heat-inactivation at 80 °C, samples were diluted 1:10 or 1:100 and assayed by *TaqMan* quantitative PCR (Li et al., *Curr. Opin. Biotechnol.* 9:43-48 (1998) on an ABI7700 Sequence Detection System (PE Biosystems, Foster City, CA) according to the manufacturer's recommendations. Fluorogenic oligonucleotides were obtained from Synthegen (Houston, TX). A standard dilution series of genomic DNA from *w*; EP(2)0750 flies was included in every experiment to allow relative quantitation of each sample. PCR primers were chosen to flank one single GATC.

Results

In vivo targeting of Dam to a specific DNA sequence: It has been demonstrated that a DNA cytosine methyltransferase can be targeted *in vitro* to a specific DNA sequence by tethering it to a DNA-binding protein (Xu et al., *Nat. Genet.* 17:376-378 (1997)). A similar approach was tested to demonstrate whether it could be used to target *E. coli* DNA adenine methyl transferase to a specific DNA locus *in vivo* in *D. melanogaster*. DNA adenine methyl transferase methylates the N^6-position of adenine in the nucleotide sequence GATC, which occurs on average every 200-300 bp in the fly genome. DNA adenine methyl transferase (DAM) was chosen because endogenous methylation of adenine does not occur in DNA of most eukaryotes. Moreover, Dam is active when expressed in yeast (Gottschling, *Proc. Natl. Acad. Sci. USA* 89:4062-4065 (1992); Singh et al., *Genes Dev.* 6:186-196 (1992); Kladde et al., *Proc. Natl. Acad. Sci. USA* 91:1361-1365 (1994)) and *Drosophila* (Wines et al., *Chromosome* 104:332-340 (1996)) and has no detectable effects on *Drosophila* development or viability (Wines et al., *Chromosoma* 104:332-340 (1996)), in contrast to certain cytosine methyltransferases (Lyko et al., *Nat. Genet.* 23:363-366 (1999)).

The well-characterized budding yeast protein GAL4 (Fischer et al., *Nature* 332:853-856 (1988)) was chosen as a DNA targeting protein. The fly line (GALDam1) was established to expresses a fusion protein (GALDam) consisting of full-length Dam and the DNA-binding domain of GAL4 ($GAL4_{DBD}$; Fig. 1A). A binding sequence for GAL4 was introduced by crossing GALDam1 flies to line EP(2)0750, which carries a P-element with 14 tandem binding sites for GAL4 (UAS_{14}) (Rorth, *Proc. Natl. Acad. Sci. USA* 93:12418-12422 (1996)) inserted into a sequenced region of chromosome 2 (Fig 1B). As a control, EP(2)0750 was crossed to the fly line Me4, which expresses Dam alone (Wines et al., *Chromosoma* 104:332-340 (1996)). The progenies of these crosses were used to test whether $GAL4_{DBD}$ was able to target Dam to GATCs in the vicinity of the UAS 14 array.

EXAMPLE 1

In this example a chromatin protein fusion protein with *E. coli* DNA adenine methyl transferase linked with Heterochromatin protein I (HP1) was used to identify DNA loci that interact with HP1 in *D. melanogaster*.

Expression vectors: The Dam open reading frame was amplified by PCR from plasmid YCpGAL-EDAM (Wines et al., *Chromosoma* 104:332-340 (1996)) and cloned into pCaSpeR-hs followed or preceded by a linker oligonucleotide encoding the *myc*-epitope tag GluGlnLysIleSerGluGluAspLeu (SEQ ID NO: 1). Resulting in vectors pNDamMyc and pCMycDam, respectively. Vector pNDamMyc carries a stop codon 15 amino acid residues after the *myc*-tag, and was used to express the Dam-myc protein. A fragment encoding amino acid residues 1-145 of GAL4 was amplified by PCR from plasmid pSPGAL1-145 (provided by S. M. Parkhurst, Fred Hutchinson Cancer Research Center, Seattle, WA) and cloned in-frame into vector pCMycDam, resulting in plasmid pGALDam. The full-length ORF of *D. melanogaster* HP1 was amplified by PCR from plasmid pTH5 (Eissenberg et al., *Proc. Natl. Acad. Sci. USA* 87:9923-9927 (1990), incorporated herein by reference) and cloned in-frame into pNDamMyc, resulting in plasmid pDamHP1.

Cell culture and immunocytochemistry: Kc167 cell culture and transfections were performed as described (Henikoff et al., *Proc. Natl. Acad. Sci. USA* 97:716-721 (2000), incorporated herein by reference). In some *in situ* staining experiments, cells were heat-shocked for 2 hours at 37 °C, followed by 5 hours recovery at 25 °C prior to fixation. *In situ* staining of proteins was carried out as described (van Steensel et al., *J. Cell. Sci.* 108:3003-3011 (1995), incorporated herein by reference) with C1A9 antibody against HP1 (James et al., *Eur. J. Cell. Biol.* 50:170-180 (1989), incorporated herein by reference), a rabbit antiserum against Cid (Henikoff et al., *Proc. Natl. Acad. Sci. USA* 97:716-721 (2000), incorporated herein by reference) or monoclonal antibody 9E10 against the *myc*-epitope tag (Santa Cruz Biotechnology, Santa Cruz, CA). For *in situ* detection of ^{m6}A in interphase cells, transfected Kc cells were grown on glass coverslips, fixed in methanol/acetic acid (3:1) for 10 minutes, washed in 70% ethanol followed by 2X SSC, denatured in 70% formamide in 2X SSC at 80°C for 10 minutes, washed in phosphate buffered saline, and stained with antibody RI 280 (Bringmann and Luhrmann, *FEBS Lett.* 213:309-315 (1987), incorporated herein by reference) following the same procedure as for proteins. Mitotic spreads were prepared by

different chromatin proteins in the same cell type can reveal functional interactions (or lack thereof) between these proteins. At the level of an organism, the profiles can be used to compare the profiles between different organisms or between different states (*e.g.*, developmental stages) of an organism. The power of the present approach is illustrated by comparative profiling of HP1 and DmSir2-1, which indicated that DmSir2-1 was not a heterochromatin protein.

Chromatin profiling can become a powerful tool in the analysis of cellular differentiation. It is anticipated that chromatin profiles made available using the methods disclosed herein for many proteins will be unique for specific cell types. Systematic mapping of such profiles can provide fundamental new insights into the mechanisms of cellular differentiation and transformation to a malignant condition. The methods as disclosed herein based on chromatin protein targeting of a nucleotide modification enzyme can be particularly useful in mammalian cells, in which other global mapping approaches based on chromatin immunoprecipitation methods (Blat and Kleckner, *Cell* 98:249-259 (1999)) may fail due to the high complexity of the genome and insufficient specificity of antibodies.

Moreover, in analogy to mRNA expression profiles (Golub et al. *Science* 286:531-537 (1999); Ross et al., *Nature Genet.* 24:227-235 (2000)), chromatin profiles can be used in studies of cellular pathology. One important application can be in the discovery and prediction of cancer types. Different classes of tumor cells are likely to display distinct chromatin profiles, and these profiles may therefore have high analytic and diagnostic value. The wide variety of chromatin proteins can allow a much more detailed and robust classification of cancer types than expression profiling, which relies on only one data set (*i.e.*, mRNA abundances) per cell type.

Methods of the present invention can also be used to provide chromatin profiles of individuals with immune deficiency or auto immune conditions as well as examining chromatin changes in reaction to various drugs and other agents. In addition chromatin binding profiles can be constructed for responses to various disease causing organisms and expression profiles can be constructed for any transcription factor of other regulatory molecule or agent.

In another embodiment of the invention, the described methods can be applied to obtaining a methylation profile. Within this method and unlike chromatin profiling which requires the introduction of a fusion protein, genomic DNA is obtained from a cell, tissue or organism of interest and from a control cell, tissue, or organism of interest.

fragments smaller than about 2.5 kb are typically added to a test array. Arrays useful in the present invention include, but are not limited to cDNA, DNA, DNA selected to contain primarily chromatin binding regions or protein binding regions, and the like. Each sample of methylated and control fractionated DNA can be labeled with, for example, a different fluorescent label. The labeled samples are mixed and applied to the array under condition conducive for hybridization using methods well known in the art. The arrays are scanned for the detection of the two labels and the loci recognized by the chromatin protein can be mapped.

Additional methods for the purification of methylated DNA regions, which can be applied separately or used in various combinations in order to further increase the purity of the isolated methylated regions include the following:

Methylated DNA fragments can be affinity purified using antibodies against ^{m6}A. Monoclonal antibody (for example, clone P1A8) which specifically recognize methyl-6-adenine (^{m6}A) have been generated using a procedure previously described (Bringmann et al., *FEBS Lett.* 213:309-315 (1987)). The antibody obtained can be used in conjunction with the restriction endonuclease *Dpn*I to affinity purify methylated DNA fragments. First, purified genomic DNA is digested with *Dpn*I, which results into exposure of ^{m6}A at the blunt ends of the digestion products. Antibody was allowed to bind to the exposed ^{m6}A. Antibody-DNA complexes were then isolated using (for example) protein A - sepharose beads (Amersham) pre-coated with rabbit-anti-mouse antibody. After purification, methylated DNA fragments were eluted from the antibody by incubation with 20 mM free methyl-6-adenosine.

Further, methylation-specific PCR amplification has been used to isolate methylated DNA fragments. After digestion with *Dpn*I, an excess of double-stranded adaptor oligonucleotide (with non-phosphorylated 5' ends to prevent self-ligation of the oligonucleotide) were ligated to the exposed blunt DNA ends using T4 DNA ligase. Because *Dpn*I cuts only methylated GATC sequences, the adaptor only ligated to methylated DNA ends. The ligated fragments were specifically amplified by PCR using a primer complementary to the adaptor sequence. The specificity of this procedure can be further enhanced using either of two modifications.

In one modification, prior to the *Dpn*I digestion, genomic DNA is treated with a DNA phosphatase such as alkaline phosphatase. This prevents ligation of the adaptor to DNA ends that were not the product of *Dpn*I digestion (*e.g.*, DNA breaks resulting from mechanical shearing r contaminant endonuclease activity during the purification of

distance of the loci recognized by the chromatin protein. It is important that the modification of a sufficient number of nucleotide residues to provide a detectable signal is not toxic to the cells, tissues or organism being tested. Therefore, as above, promoters which provide for low levels of expression are used and nucleotide modification enzymes which provide non-toxic nucleotide modifications are used.

Detection of Chromatin Binding Sites:

Several methods are available for the detection of modified nucleotides in the vicinity of the binding loci recognized by the chromatin protein. These include, but are not limited to, immunohistochemistry, Southern blot analysis, PCR analysis and array (*i.e.*, macro- and micro-array) analysis.

In a typical embodiment, cells are grown or collected on a solid phase appropriate for microscopy. For example, transformed cells can be cultured on a glass microscope cover slip. The cells are then fixed and washed. An antibody specific for the nucleotide modification carried out by the nucleotide modification enzyme of the fusion protein is added. The antibody can be either polyclonal antisera or a monoclonal antibody. Antibody can be labeled directly or a second labeled antibody can be used to detect the nucleotide modification. Following an incubation period the cells are washed and the antibody is detected providing a location within the nucleus where the chromatin protein complexes within the chromatin. In one particular embodiment the cells are prepared as mitotic spreads by methods well known to the skilled artisan.

A wide variety of labels can be employed for detection of the nucleotide modification. For example, the label can be, chemiluminescent, enzyme, fluorophor, or a radioactive moiety, and the like. Typically, fluorescent labels, such as, fluorescein, phycoerythrin (PE), Cy3, Cy5, Cy7, Texas Red, allophycocyanin (APC), Cy7APC, Cascade Blue, Cascade Yellow, and the like, can be used. Methods for labeling antibodies are well known to the skilled artisan.

In still another embodiment Southern blot can be used to map the region of the chromatin where a nucleotide modification has occurred. Typically, genomic DNA is isolated from a population of cells transformed with the vector capable of expressing the chromatin binding protein-nucleotide modification enzyme fusion polypeptide by methods well known to the skilled artisan. The population of cells useful in the methods of the present invention can be isolated from cells grown *in vitro*, isolated from a single tissue, or isolated from a multicellular organism.

Nucleotide modifying enzymes, fragments, derivatives and analogs thereof useful in the present invention are those which can modify one or more nucleotides in a nucleic acid sequence, such as an RNA, DNA, or the like, under conditions found in a live cell and in a manner which is detectable. The enzyme must also modify the nucleotides in a manner which is not toxic to the cell. In other words, the cell or organism must be able to continue to proliferate and differentiate in a normal manner. For the modification to be detectable, an enzyme is selected which modifies the nucleotide in a manner which is not typical of a modification commonly found in the cell being assayed. For instance, in eukaryotic cells it is typical to select as the modification enzyme, for example, DNA adenine methyl transferase because methylation of adenine is not common in eukaryotic cells. Additional nucleotide modification enzymes useful in the present invention include, for example, but are not limited to, adenine methyltransferases, cytosine methyltransferases, thymidine hydroxylases, hydroxymethyluracil β-glucosyl transferases, adenosine deaminases, and the like. However, as described in more detail below, within one embodiment, a modification of the method of the present invention relies on an endogenous modification enzyme to modify DNA in a cell, the sites of such modifications are then determined by a variety of detection means, including the use of nucleic acid arrays.

In the methods of the present invention, the DNA modification enzyme, fragment, derivative, or analog thereof, is targeted to the loci associated with the binding of the chromatin protein by the chromatin protein, fragment, derivative or analog thereof, as a fusion protein. Typically, the polypeptides which comprise the chromatin protein and the DNA modification enzyme are separated from one another by one or more amino acid residues which comprise a linker sequence. The linker can be from about 1 to about 1000 amino acid residues, or more. Typically, the linker sequence is from about 3 to about 300 amino acid residues. The amino acid sequence can be from another polypeptide or can be an artificial sequence of amino acid residues, such as, for example, Gly and Ser residues which provide a flexible linear amino acid sequence allowing the amino acid sequences for the chromatin polypeptide and the nucleotide modification enzyme to fold into an active configuration. In a particular embodiment of the present invention a linker peptide comprising the myc-epitope tag GluGlnLysIleSerGluGluAspLeu (SEQ ID NO: 1) was inserted between the chromatin polypeptide and the nucleotide modification enzyme DNA adenine methyl transferase.

predominantly to pericentric genes and transposable elements, GAGA factor (GF) which associates with euchromatic genes that are enriched in $(GA)_n$ motifs, or a Drosophila homolog of the yeast *Sir2* gene (*DmSir2-1*) which associates with certain active genes can be used to construct a fusion protein of the invention. Fragments, derivatives or analogs of a chromatin protein or protein complex can be tested for the desired activity by procedures known in the art, including but not limited to the functional assays to determine whether the fragment recognizes and binds the target loci or nucleotide sequence recognized by the native full length chromatin binding protein. The affinity or avidity of the binding to the target loci or nucleotide sequence can be the same, less or greater than the affinity or avidity of the native full length protein. It is only necessary that the fragment, derivative or analog recognize and bind the target loci or sequence. In addition, the chromatin polypeptide fragment, derivative, or analog can be tested for the desired activity in the fusion protein to ensure localization to the appropriate loci.

Polypeptide derivatives include naturally-occurring amino acid sequence variants as well as those altered by substitution, addition or deletion of one or more amino acid residues that provide for functionally active molecules. Polypeptide derivatives include, but are not limited to, those containing as a primary amino acid sequence all or part of the amino acid sequence of a native chromatin polypeptide including altered sequences in which one or more functionally equivalent amino acid residues (*e.g.*, a conservative substitution) are substituted for residues within the sequence, resulting in a silent change.

In another aspect, polypeptides of the present invention include those peptides having one or more consensus amino acid sequences shared by all members of the chromatin protein family members, but not found in other proteins. Database analysis indicates that these consensus sequences are not found in other polypeptides, and therefore this evolutionary conservation reflects the nucleotide target binding-specific function of chromatin polypeptides. Chromatin polypeptide family members, including fragments, derivatives and/or analogs comprising one or more of these consensus sequences, are also within the scope of the invention.

In another aspect, a polypeptide consisting of or comprising a fragment of a chromatin polypeptide having at least 5 contiguous amino acids of the chromatin polypeptide which recognize the specific target nucleotide sequence is provided. In other embodiments, the fragment consists of at least 20 or 50 contiguous amino acids of the chromatin polypeptide. In a specific embodiment, the fragments are not larger than 35, 100 or even 200 amino acids.

aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.* (1990), *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits in each direction is halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLAST program uses as defaults a word length (W) of 11, the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915-9 (1992), which is incorporated by reference herein) alignments (B) of 50, expectation (E) of 10, M=5, N=-4, and a comparison of both strands.

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, *e.g.*, Karlin & Altschul, *Proc. Natl. Acad. Sci. USA* 90:5873-77 (1993), which is incorporated by reference herein). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more typically less than about 0.01, and most typically less than about 0.001. Further, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

The terms "transformation" or "transfection" means a process of stably or transiently altering the genotype of a recipient cell or microorganism by the introduction of polynucleotides. This is typically detected by a change in the phenotype of the recipient cell or organism. The term "transformation" is generally applied to microorganisms, while "transfection" is used to describe this process in cells derived from multicellular organisms.

Generally, other nomenclature used herein and many of the laboratory procedures in cell culture, molecular genetics and nucleic acid chemistry and hybridization, which are described below, are those well known and commonly employed in the art. (See generally Ausubel *et al.* (1996) *supra*; Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor Laboratory Press, New York (1989), which are

The term "substantial similarity" in the context of polypeptide sequences, indicates that the polypeptide comprises a sequence with at least 70% sequence identity to a reference sequence, or preferably 80%, or more preferably 85% sequence identity to the reference sequence, or most preferably 90% identity over a comparison window of about 10-20 amino acid residues. In the context of amino acid sequences, "substantial similarity" further includes conservative substitutions of amino acids. Thus, a polypeptide is substantially similar to a second polypeptide, for example, where the two peptides differ only by one or more conservative substitutions.

The term "conservative substitution," when describing a polypeptide, refers to a change in the amino acid composition of the polypeptide that does not substantially alter the polypeptide's activity. Thus, a "conservative substitution" of a particular amino acid sequence refers to substitution of those amino acids that are not critical for polypeptide activity or substitution of amino acids with other amino acids having similar properties (*e.g.*, acidic, basic, positively or negatively charged, polar or non-polar, and the like) such that the substitution of even critical amino acids does not substantially alter activity. Conservative substitution tables providing functionally similar amino acids are well known in the art. For example, the following six groups each contain amino acids that are conservative substitutions for one another: 1) alanine (A), serine (S), threonine (T); 2) aspartic acid (D), glutamic acid (E); 3) asparagine (N), glutamine (Q); 4) arginine (R), lysine (K); 5) isoleucine (I), leucine (L), methionine (M), valine (V); and 6) phenylalanine (F), tyrosine (Y), tryptophan (W). (*See also* Creighton, *Proteins*, W. H. Freeman and Company (1984).) In addition, individual substitutions, deletions or additions that alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservative substitutions."

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, for example, by the local homology algorithm of Smith & Waterman (*Adv. Appl. Math.* 2:482 (1981), which is incorporated by reference herein), by the homology alignment algorithm of Needleman & Wunsch (*J. Mol. Biol.* 48:443-53 (1970), which is incorporated by reference

"Chromatin protein" includes, but is not limited to histones, transcriptional factors, centromere proteins, heterochromatin proteins, euchromatin proteins, condensins, cohesins, origin recognition complexes, histone kinases, dephosphorylases, acetyltransferases, deacetylases, methyltransferases, demethylases, and other enzymes that covalently modify histone, DNA repair proteins, proteins involved in DNA replication, proteins involved in transcription, proteins part of dosage compensation complexes and X-chromosome inactivation, proteins that are part of chromatin remodeling complexes, telomeric proteins, and the like.

"Chromatin protein-enzyme fusion polypeptide" refers to a polypeptide encoded by a polynucleotide encoding the chromatin protein operatively associated with a polynucleotide which encodes a nucleotide modification enzyme. Also encompassed within this definition are polynucleotides which encode a functionally active fragment, derivative or analog of the chromatin protein or nucleotide modification enzyme. The term "polypeptide" refers to a polymer of amino acids and its equivalent and does not refer to a specific length of the product; thus, peptides, oligopeptides and proteins are included within the definition of a polypeptide. A "fragment" refers to a portion of a polypeptide having typically at least 10 contiguous amino acids, more typically at least 20, still more typically at least 50 contiguous amino acids of the chromatin protein. A "derivative" is a polypeptide which is identical or shares a defined percent identity with the wild-type chromatin protein or nucleotide modification enzyme. The derivative can have conservative amino acid substitutions, as compared with another sequence. Derivatives further include, for example, glycosylations, acetylations, phosphorylations, and the like. Further included within the definition of "polypeptide" are, for example, polypeptides containing one or more analogs of an amino acid (*e.g.*, unnatural amino acids, and the like), polypeptides with substituted linkages as well as other modifications known in the art, both naturally and non-naturally occurring. Ordinarily, such polypeptides will be at least about 50% identical to the native chromatin binding protein or nucleotide modification enzyme acid sequence, typically in excess of about 90%, and more typically at least about 95% identical. The polypeptide can also be substantially identical as long as the fragment, derivative or analog displays similar functional activity and specificity as the wild-type chromatin protein or nucleotide modification enzyme.

The terms "amino acid" or "amino acid residue", as used herein, refer to naturally occurring L amino acids or to D amino acids as described further below. The commonly used one- and three-letter abbreviations for amino acids are used herein (*see, e.g.*,

methylation in Dam-HP1 transfected cells and Dam-myc transfected cells, calculated from the data in Fig. 3B. Shading is the same as the boxes in Fig. 3A. Bullets indicate ratios that are significantly different (one, $p<0.05$; two, $p<0.01$; three, $p<0.001$ according to the Mann-Whitney U-test) from the pooled ratios of the four heterochromatic loci (black bullets) or the five euchromatic loci (white bullets). Error bars represent standard deviations. The number of observations is indicated in parentheses.

Fig. 4A and Fig. 4B depict mapping of HP1 target loci. Fig. 4A demonstrates a chromosomal map of Cy3:Cy5 ratios (representative experiment). Probed loci are indicated by their approximate position on the cytogenetic map. Centromeres are indicated by ovals. The large heterochromatic proximal region of the X chromosome is depicted as a rectangle to the left of the centromere (not to scale). Some genes with relatively high levels of HP1 binding are labeled. Fig. 4B depicts dispersed repetitive elements (mostly transposons).

Fig. 5A through Fig. 5C depict the mapping of GAF target loci. Fig. 5A provides a chromosomal map of Cy3:Cy5 ratios (average of two experiments) using a GAF-Dam fusion protein. Some genes with relatively high levels of GAF binding are labeled. Fig 5B depicts dispersed repetitive elements (mostly transposons). Fig. 5C depicts a box plot showing the relative abundances of GAGAG (SEQ ID NO: 4) and GAGAGAG (SEQ ID NO: 5) sequence elements in probed regions with low (open boxes) and high (filled boxes) levels of GAF binding. Horizontal lines represent the 10th, 25th, 50th (median), 75th and 90th percentiles. p-values are according to the Mann-Whitney U-test.

Fig. 6A and Fig. 6B depict the mapping of DmSir2-1 target loci. Fig. 6A provides a chromosomal map of Cy3:Cy5 ratios (average of two experiments) for chromosomes 2, 3 and 4, and the X chromosome. Fig. 6B depicts dispersed repetitive elements (mostly transposons). Some genes of particular interest or with high levels of DmSir2-1 are labeled.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

The terms "polynucleotide" and "nucleic acid" refer to a polymer composed of a multiplicity of nucleotide units (ribonucleotide or deoxyribonucleotide or related structural variants) linked via phosphodiester bonds. A polynucleotide or nucleic acid can be of substantially any length, typically from about six (6) nucleotides to about 10^9 nucleotides or larger. Polynucleotides and nucleic acids include RNA, cDNA, genomic

embodiment of the present invention a polynucleotide encoding *Escherichia coli* DNA adenine methyltransferase was used as the tethered nucleotide modification enzyme.

Once the nucleotide modification enzyme has been directed to the chromatin binding site by the chromatin protein, the nucleotide modification enzyme can modify nucleotides of the chromatin in the region of the binding site. These modifications of the nucleotides can be detected by various methods including immunochemistry, Southern blot, PCR, and various types of macro- and micro-arrays. The binding loci of the chromatin protein can be identified by determining the location of the detected nucleotide modifications within the chromatin. In a specific embodiment, the loci or the chromatin proteins heterochromatin binding protein 1, GAGA factor and Drosophila *DmSir2-1* gene was determined by immunocytochemistry.

The methods of the present invention also provide methods for large scale mapping of loci of chromatin proteins. The methods can be used to obtain detailed genome-wide maps of the binding patterns of chromatin proteins in, for example, cell populations grown in culture, tissues, or in cells isolated from an entire multicellular organism. The chromatin profiles can provide information into the functions and mechanisms of action of chromatin proteins on an individual cellular basis, at the tissue level, and the organism level. In a particular embodiment pairwise comparison of profiles of different chromatin proteins in the same cell type can be used to determine functional interactions (or lack thereof) between these proteins. At the level of an organism, the profiles can be used to compare the profiles between different organisms or between different states (*e.g.*, developmental stages) of an organism.

The present invention further provides methods for producing a profile of chromatin protein loci for a cell population of interest which method comprises; transfecting the cell population with a plurality of expression vectors capable of expressing a plurality of different chromatin protein-nucleotide modification enzyme fusion proteins, each expression vector comprising a nucleic acid encoding a low efficiency promoter operatively associated with a nucleic acid encoding the different chromatin proteins and a nucleic acid encoding a nucleotide modification enzyme; culturing the transfected cells for a period of time sufficient for expression of and binding of each of the plurality of chromatin protein-nucleotide modification enzyme fusion polypeptides; and detecting the loci for each of the nucleotide modifications within the chromatin of the cell population. The profile of chromatin protein loci for the cell population is determined from the location of the DNA modifications.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 September 2001 (20.09.2001)

PCT

(10) International Publication Number
WO 01/68807 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US01/08590

(22) International Filing Date: 16 March 2001 (16.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/190,362 16 March 2000 (16.03.2000) US
60/ 1 March 2001 (01.03.2001) US

(71) Applicant *(for all designated States except US)*: **FRED HUTCHINSON CANCER RESEARCH CENTER** [US/US]; Office of Technology Transfer, 1100 Fairview Avenue North, M/S: C2M 027, Seattle, WA 98109-1024 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **VAN STEENSEL, Bas** [NL/NL]; Nieuwegrachtje 1-3, NL-1011 VP Amsterdam (NL). **HENIKOFF, Steven** [US/US]; 4711 51st Place SW, Seattle, WA 98116 (US).

(74) Agents: **POOR, Brian, W.** et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111 (US).

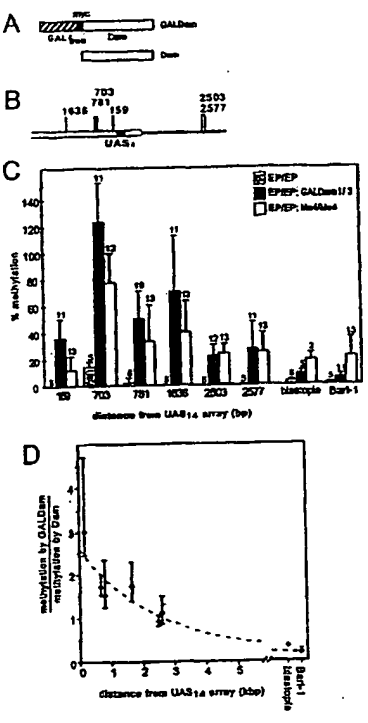
(81) Designated States *(national)*: AU, CA, JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: IDENTIFICATION OF *IN VIVO* DNA BINDING LOCI OF CHROMATIN PROTEINS USING A TETHERED NUCLEOTIDE MODIFICATION ENZYME



(57) Abstract: A novel technique is provided, designated DamID, for the identification of DNA loci that interact *in vivo* with specific nuclear proteins in eukaryotes. By tethering a DNA modification enzyme, in particular, *E. coli* DNA adenine methyl transferase (Dam), to a chromatin protein. The DNA modification enzyme (Dam) can be targeted *in vivo* to the native binding loci of the protein, resulting in local DNA modification. Sites of DNA modification can subsequently be mapped using modification-specific restriction enzymes, antibodies, or DNA array methods. DNA Modification Identification (DamID) has potential for genome-wide mapping of *in vivo* target binding sites of chromatin proteins in various eukaryotes.

WO 01/68807 A2

IDENTIFICATION OF *IN VIVO* DNA BINDING LOCI OF CHROMATIN PROTEINS USING A TETHERED NUCLEOTIDE MODIFICATION ENZYME

RELATED APPLICATIONS

This application is a continuation in part of United States patent application Serial number 60/_____, filed March 1, 2001, and a continuation in part of United States patent application 60/190,362, filed March 16, 2000, the disclosures of which are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

Chromatin is the highly complex structure consisting of DNA and hundreds of directly and indirectly associated proteins. Most chromatin proteins exert their regulatory and structural functions by binding to specific chromosomal loci. Knowledge of the nature of the *in vivo* target loci is essential for the understanding of the functions and mechanisms of action of chromatin proteins. Interactions between protein complexes and DNA are at the heart of essential cellular processes such as transcription, DNA replication, chromosome segregation, and genome maintenance. High-resolution, genome-wide maps of binding sites of these proteins can provide a valuable resource for researchers studying chromosome organization, chromatin structure, and gene regulation, but such comprehensive maps are currently unavailable. Therefore, techniques are needed to identify DNA loci that interact *in vivo* with specific proteins.

At present, only a few techniques are available to localize the genomic loci recognized by DNA binding proteins (reviewed in Simpson, *Curr. Opin. Genet. Dev.* 9:225-229 (1999)). *In situ* cross-linking methods followed by the immunoprecipitation purification of protein-DNA complexes have been used to test the interaction of individual chromosomal loci with a particular chromatin protein (Solomon et al., *Cell* 53:937-947 (1988); Law et al., *Nucleic Acids Res.* 26:919-924 (1988); Orlando et al., *Methods* 11:205-214 (1997); Kuo and Allis, *Methods* 19:425-433 (1999); Blat et al., *Cell* 98:249-259 (1999); Orlando, *Trends Biochem. Sci.* 25:99-104 (2000)). These previously disclosed techniques have the inherent risk of artifacts induced by the cross-linking reagent, and highly specific antibodies against each protein of interest are required, as well as relatively large numbers of cells. A modification of this approach was recently employed to identify binding sites of cohesins along a complete chromosome in yeast (Blat and Kleckner, *Cell* 98:249-259 (2000).